MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2022/23

ECE3226 - ADVANCED COMPUTER ARCHITECTURE AND PARALLEL COMPUTING

(All sections / Groups)

12 JULY 2023
9:00 a.m. – 11:00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 5 pages including cover page and appendices with 4 Questions only.

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the answer booklet.

Question 1

a) Flynn taxonomy is a classification system for parallel computer architectures, proposed by computer scientist Michael J. Flynn in 1966. The taxonomy categorizes computer architectures into four main classes based on their instruction and data streams: Single Instruction Single Data (SISD), Single Instruction Multiple Data (SIMD), Multiple Instruction Single Data (MISD), and Multiple Instruction Multiple Data (MIMD). Define and elaborate the general characteristics of SIMD computer with the aid of a sketch.

[8 marks]

b) In GPU-accelerated applications, the sequential part of the workload is running on the CPU (which is optimized for single-threaded performance), while the compute intensive portion of the application is running on thousands of GPU processing cores in parallel. Explain in your own words, on how the used of registers and shared memory that is residing on a GPU chip could relate to the acceleration of compute-intensive portions of an application program.

[4 marks]

c) Direct memory access (DMA) is a method that allows an input/output (I/O) device to send or receive data directly to or from the main memory, bypassing the CPU to speed up memory operations Specify three types of DMA transfer techniques available.

[3 marks]

d) Analyse the impact of increasing the clock rate on the design of a new computer (computer B) to speed up a program (P1) compared to an existing computer (computer A). The clock oscillator of computer A is 8GHz and program P1 takes 10 second to run on computer A. Computer B needs to run program P1 in 6 seconds. To achieve this, the clock rate needs to be increased, but this will affect the rest of processing unit design, causing computer B to require 1.5 times as many clock cycles as computer A for this program. Based on this information:

i) Estimate the number of CPU clock cycles for computer A, and

[5 marks]

ii) Estimate the clock rate for computer B.

[5 marks]

Continued ...

Question 2

(a) Symmetric multiprocessing or shared-memory multiprocessing (SMP) involves a type of computer system where multiple processors are connected to a single, shared main memory and controlled by a single operating system. One type of SMP architecture is called "Shared Bus" architecture, where all processors and devices are connected to a single shared communication pathway. Elaborate and describe two advantages and two disadvantages of using the Shared Bus architecture for symmetric multiprocessing.

[6 marks]

(b) Pipelining is a technique of decomposing a sequential process into smaller sub-operations, each of which is performed by a dedicated segment of hardware. However, certain factors can disrupt the normal performance of pipelining, causing delays and other issues. Briefly describe five factors that can cause pipelining to deviate from its normal performance. [5 marks]

(c) When a computer uses pipelining to speed up processing, sometimes under certain circumstances, (i.e. pipeline hazards) can arise which cause delays or prevent the execution of the next instruction in the pipeline. Elaborate THREE different types of hazards that can occur in a pipelined computer, and explain what each one does to impede the normal flow of instructions. [6 marks]

(d) The architecture of DLX was chosen based on observations about most frequently used primitives in programs. DLX provides a good architectural model for study on how to balance the length of each pipeline stage. List FIVE phases of DLX operations.

[5 marks]

(e) Superscalar processors were first developed in the 1980s to improve the overall performance of the processor to allow processor to perform more work per clock cycle compared to a traditional scalar processor. Whereas a very long instruction word (VLIW) processor enables multiple operation in a single long instruction word that contains several operations to achieve high instruction throughput. List THREE comparisons among superscalar and VLIW.

[3 marks]

Continued ...

Question 3

(a) Vector processor is basically a central processing unit that can execute operations on multiple data elements, or vectors, in a single instruction. Explain the concepts of initiation rate, convoy, chime and vector start- up time as they relate to vector processors.

[8 marks]

(b) Provide THREE crucial factors that can affect the execution time of a vector processor. [3 marks]

(c) As a computer engineer, you are responsible for optimizing the workflow of a new microprocessor, which will be executing a workload program core consisting of 10^6 instructions, each taking 200 picoseconds to complete. Analyze the impact of pipelining on the performance of this program core, taking into account both the benefits and potential overhead of pipelining.

i) Estimate the time required to execute the program core on a non-pipelined processor. [2 marks]

ii) Assuming the current microprocessor implements an architecture with 15 pipeline stages, estimate the maximum speedup achievable compared to a non-pipelined processor by assuming it is perfectly pipelined. [2 marks]

iii) During the implementation of pipelining in practical situations, overhead tends to occur for every pipeline stage. Analyse the effects of this overhead on instruction latency and instruction throughput, and discuss the design considerations necessary to minimize these effects.

[2 marks]

(d) Given a bank busy time takes 6 clock cycles and a total memory latency is estimated to be 12 cycles, if we are given 8 memory banks:

i. Calculate the number of clock cycles required to complete a 64 element vector load with a stride of 1. [3 marks]

ii. Estimate the number of clock cycles required to complete a 64 element vector load with a stride of 32 [3 marks]

(e) Show the following code sequence lays out in convoys, assuming a single copy of each vector functional unit:

```
LV      V1,Rx      ;load vector X
MULVS.D V2,V1,F0   ;vector-scalar multiply
LV      V3,Ry      ;load vector Y
ADDVV.D V4,V2,V3   ;add two vectors
SV      V4,Ry      ;store the sum
```

[2 marks]

Continued ...

Question 4

(a) Micro-coded functions can improve the performance of a microprocessor, but it will add a layer of complexity. Microcode can be characterized as horizontal or vertical. Define the meaning of microcode and elaborate the differences between horizontal microcode and vertical microcode.

[7 marks]

(b) Cache coherency is a situation where multiple processor cores share the same memory hierarchy, but each processor has its own "quick access" storage area (known as a cache). In a cache coherent system, the caches are responsible for making sure that everyone is working with the same information. Briefly describe what is migration and replication for enforcing coherence.

[2 marks]

(c) The Amdahl's Law inequality is sometimes treated as a performance model, instead of a performance limit, list down THREE problems faced by Amdahl's law in parallel computer?

[3 marks]

(d) Construct a data-flow graph for the following nested if-then-else statement.

```
if x > y then
        if k = i then
          z = a + b
        else
          z = c * d
else
        z = e + f
endif
```

[13 marks]

End of Paper